SCIENTIFIC METHODS AND TECHNOLOGIES

UDC 519.233.33

Chekhlystova Yu.A., Elshin V.R., Glazkova M.Yu. Testing the distribution hypothesis. Pearson's Criterion

Проверка гипотезы о распределении. Критерий Пирсона.

Chehlystova Julia Alexandrovna
2nd year student of the specialty Construction of unique buildings and structures (SUZ-233)
Faculty of Civil Engineering,
Voronezh State Technical University
Yelshin Vladislav Romanovich
2nd year student of the specialty Construction of unique buildings and structures (SUZ-232)
Faculty of Civil Engineering,
Voronezh State Technical University
Glazkova Maria Yurievna
Ph.D, Associate Professor
Department of Applied Mathematics and Mechanics
Voronezh State Technical University
Чехлыстова Юлия Александровна
студентка 2 курса специальности Строительство уникальных зданий и сооружений (СУЗ-233)
строительный факультет, Воронежский государственный технический университет
Ельшин Владислав Романович
студент 2 курса специальности Строительство уникальных зданий и сооружений (СУЗ-232)
строительный факультет, Воронежский государственный технический университет
Глазкова Мария Юрьевна
к. ф.-м. н., доцент кафедры прикладной математики и механики
Воронежский государственный технический университет

Аннотация. *Проверка статистических гипотез помогает оценить, даст ли необходимый результат реализация некоторых планов или нет, говоря о любой сфере: бизнес, наука, учеба, медицина. Критерий Пирсона нужен для проверки гипотез о распределении, он позволяет:*

1. *Выяснить, существует ли связь между некоторыми переменными;*

2. *Оценить значимость различий между фактическим количеством исходов и теоретическим.*

Ключевые слова: *гипотеза, критерий согласия Пирсона, уровень значимости, критическая область, выборочные данные, функция распределения, число степеней свободы, случайная величина X.*

Abstract. *Checking statistical hypotheses helps to assess whether the implementation of certain plans will give the necessary result or not, speaking about ideas in any field: business, science, education, medicine. The Pearson criterion is needed to test hypotheses about the distribution, it allows:*

1. Find out if there is a relationship between some variables;

2. To assess the significance of the differences between the actual number of outcomes and the theoretical one.

Keywords: *hypothesis, Pearson's criterion of agreement, significance level, critical area, sample data, distribution function, number of degrees of freedom, random variable.*

Рецензент: Мартеха Александр Николаевич – кандидат технических наук, доцент.
Доцент ФГБОУ ВО «РГАУ-МСХА им. К.А. Тимирязева»

В работе изложено определние и методика проверки гипотезы о распредлении с помощью критерия Пирсона, а также приведены практические примеры по решению задач данным методом.

Целью работы является изучение методики проверки статистических гипотез о распредлении с помощью критерия Пирсона.

В работе будет рассмотрено понятие статистической гипотезы о распределении, изучен порядок проверки статистической гипотезы с помощью критерия Пирсона и решены математические задачи с помощью проверки гипотез по критерию Пирсона, в которых предпололагаются различные теоритические законы распределения.

Выборка наблюдений случайной величины – это последовательность независимых случайных величин, которые соответствуют всем возможным результатам некоторого количества статистических экспериментов и имеют один закон распределения вероятностей со случайной величиной.

Функция распределения случайной величины – вероятность того, что случайная величина X примет значение меньшее, чем некоторое заданное значение X.

Нулевая гипотеза – утверждение, которое делается с целью проверки статистических гипотез.

Критическая область – это совокупность значений некоторого критерия, при которых нулевая гипотеза отвергается.

Уровень значимости – вероятность отвергнуть верную гипотезу, максимально приемлемый для учёного риск получения ложноположительного результата.

Число степеней свободы – это количество значений, используемых при расчетах статистических характеристик, которые могут свободно изменяться.

Статистическая гипотеза — это предположение о характеристиках, свойствах, параметрах объектов исследования, генеральных совокупностей в целом и их отдельных компонентов.

В ходе проверки статистических гипотез о соответствии отдельных параметров закона распределения случайных величин предполагалось, что законы распределения этих величин известны. Однако при решении практических задач модель закона распределения в общем случае заранее неизвестна, поэтому возникает необходимость выбора модели закона распределения, согласующейся с результатами выборочных наблюдений.

Пусть $x_1, x_{2,\ \ldots,}\ x_n$ - выборка наблюдений случайной величины X с неизвестной непрерывной функцией распределения F(x). Проверяется гипотеза H_0, утверждающая, что X распределена по закону, имеющему функцию распределения F(x), равную функции $F_0(x)$, т.е. проверяется нулевая гипотеза H_0: F(x)= $F_0(x)$. Критерии, используемые для проверки нулевой гипотезы о неизвестном распределении, называются критериями согласия. Рассмотрим критерий согласия Пирсона.

Схема проверки нулевой гипотезы

1. По выборке $x_1, x_{2,\ \ldots,}\ x_n$ строят вариационный ряд; он может быть как дискретным, так и интервальным. Для определенности рассмотрим дискретный вариационный ряд.
2. По предварительным данным или по данным предыдущих исследований делают предположение (принимают гипотезу) о модели закона распределения случайной величины X.
3. По выборочным данным проводят оценку параметров выбранной модели закона распределения. Считаем. что закон распределения имеет r параметров (например, нормальный – два параметра (a_0, σ_0), биноминальный закон имеет один параметр p и т.д.)
4. Подставляя выборочные оценки значений параметров распределения, находят теоретические значения вероятностей $p_i^T = P(X = x_i),\ i - 1{,}2,\ \ldots,\ k$
5. По выборке $x_1, x_{2,\ \ldots,}\ x_n$ строят вариационный ряд; он может быть как дискретным, так и интервальным. Для определенности рассмотрим дискретный вариационный ряд: $m_i^T = p_i^T$n, где n=$\sum_{i=1}^{k} m_i$
6. Рассчитывают значение критерия согласия Пирсона:

$$\chi_r^2 = \sum_{i=1}^{k} \frac{(m_i - m_i^T)^2}{m_i^T} \quad (1)$$

Эта величина при $n \to \infty$ стремится к распределению χ^2 с $l = k - r - 1$ степенями свободы. Поэтому для расчетов используют таблицы распределения χ^2.

7. Зная уровень значимости α, находят критическую область (она всегда правосторонняя) $((\chi^2_{\text{кр}})^{\Pi}; \infty)$; значение $(\chi^2_{\text{кр}})^{\text{п}}$ определяют из соотношения $\alpha = P(\chi^2 > (\chi^2_{\text{кр}})^{\text{п}})$.

При попадании численного значения χ_r^2 в интервал $((\chi^2_{\text{кр}})^{\Pi}; \infty)$, гипотеза H_0: $F(x) = F_0(x)$ отклоняется и принимается альтернативная гипотеза о том, что выбранная модель закона распределения не подтверждается выборочными данными, при этом допускается ошибка, вероятность которой равна α.

Рассмотрим следующую задачу:

Фирма по изготовлению лекарств «Здоровье» выпустила 10 новых видов препаратов для иммунной системы. Было опрошено 300 посетителей данной аптеки и выявлено, сколько каждому потребовалось препаратов данной фирмы в течение 3 месяцев. Результаты опроса представлены в таблице 1.

Таблица 1

i	1	2	3	4	5	6	7	8	9	10	11
x_i	0	1	2	3	4	5	6	7	8	9	10
m_i	55	56	64	22	27	17	20	12	11	13	3

Проверить, подчинена ли случайная величина X биномиальному закону распределения?

Предположим, вероятность покупки одного лекарства не зависит от решения о необходимости покупки других типов лекарств.

Вероятность купить любой конкретный тип лекарств одна и та же и равна q=1-p.

С учетом вышесказанного предположим, что X подчинена биноминальному закону распределения (нулевая гипотеза H_0) , т.е. вероятность того, что покупатель приобретет x товаров, может быть посчитана по формуле 2:

$$P(X = x) = C_{10}^x p^x q^{10-x} \quad (2)$$

Найдем среднее число товара, купленного одним посетителем аптеки по формуле 3:

$$\bar{x} = \frac{\sum_{i=1}^{11} x_i m_i}{\sum_{i=1}^{11} m_i} \quad (3)$$

Р - это вероятность того, что покупатель купить товар. Оценкой вероятности р является относительная частота p^*, которая вычисляется по формуле 4:

$$p^* = \frac{\bar{x}}{\nu} \quad (4)$$

Где ν-число товаров, из которых может выбрать каждый покупатель

Таким образом по формулам 3-4:

$p^* = \frac{\bar{x}}{\nu} = \frac{\sum_{i=1}^{11} x_i m_i}{\nu \sum_{i=1}^{11} m_i}$=(0·0,18+1·0,19+…+10·0,01)=0,294

Подставим значения p^* =0,294 и q^*= 1-0,294=0,706 в выражение (1) и при различных x_i получим теоретические вероятности p_i^T (по формуле 2) и частоты $m_i^T = p_i^T n$ (таблица 2).

Таблица 2

Промежуточные данные

Номер группы i	x_i	p_i^T	m_i^T
1	0	0,030764298	9,229289
2	1	0,128111949	38,43358
3	2	0,240073809	72,02214
4	3	0,266597544	79,97926
5	4	0,194283904	58,28517
6	5	0,097086914	29,12607
7	6	0,033691635	10,10749
8	7	0,008017273	2,405182
9	8	0,001251989	0,375597
10	9	0,000115859	0,034758
11	10	4,82473E-06	0,001447

Из таблицы 1 видно, что для групп 8-11 теоретическая частота $m_i^T < 5$. Объединим эти группы с соседними, результаты занесем в таблицу 3.

Таблица 3

Промежуточные данные

Номер группы i	x_i	m_i	m_i^T
1	0	55	9,229289
2	1	56	38,43358
3	2	64	72,02214
4	3	22	79,97926
5	4	27	58,28517
6	5	17	29,12607
7	6	20	10,10749
8	7-10	39	2,816984

По данным таблицы 2 и формуле 1 рассчитываем величину критерия согласия:

$$\chi^2 = \frac{(55-9{,}23)^2}{9{,}23} + \frac{(56-38{,}43)^2}{38{,}43} + \cdots + \frac{(39-2{,}82)^2}{2{,}82} = \\ = 774{,}2$$

При уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $l = k - r - 1 = 7 - 1 - 1 = 5$ по таблице $\left(\chi^2_{kp}\right)^{\Pi} = 11{,}1$.

Величина $\chi^2_r = 774{,}2\, \epsilon\, (11{,}1; \infty)$, значит, нулевая гипотеза должна быть отвергнута.

Рассмотрим еще одну задачу:

При производстве булочек, рассчитывали, что в день в каждой из 5 кондитерских, расположенных в различных районах, объем продаж будет одинаковый. В действительности объем продаж в кондитерских оказался иным (таблица 4).

Таблица 4

Объем продаж в кондитерских

Район	i	1	2	3	4	5
Фактический объем продаж	m_i	106	116	82	110	86

Определить, значимы или нет различия между фактическими и теоретическими объемами продаж, считая уровень значимости равным 0,01 и 0,1.

Так как в задаче спрашивается о согласовании ожидаемых (одинаковых) и фактических объемов продаж, то теоретический «закон распределения» определен: во всех районах объем продаж одинаков, т.е.

$$m_1^T = m_2^T = m_3^T = m_4^T = m_5^T = \frac{\sum_{i=1}^{5} m_i}{5} = \frac{500}{5} = 100$$

Заметим, что в данном примере нельзя использовать в качестве закона распределения биноминальный или нормальный закон, так как речь идет об одновременном сравнении пяти районов.

Тогда рассчитывая значение критерия согласия Пирсона по формуле 1:

$$\chi_r^2 = \sum_{i=1}^{5} \frac{\left(m_i - m_i^T\right)^2}{m_i^T} = \frac{1}{100}(36 + 256 + 324 + 100 + 196) = 9{,}12$$

Выбирая уровень значимости $\alpha = 0{,}01$, по таблице 7 для числа степеней свободы $l = 5 - 1 = 4$ находим $\left(\chi^2_{\text{кр}}\right)^{\Pi}$=13,3, а для уровня значимости $\alpha = 0{,}05$ при $l = 4$, соответственно, $\left(\chi^2_{\text{кр}}\right)^{\Pi}$=9,5.

Следовательно, для уровня значимости $\alpha = 0{,}01$ критическая область представляет собой интервал $(13{,}3; \infty)$, $\chi_r^2 = 9{,}12$ не попадает в критическую область, т.е. нулевая гипотеза, состоящая в том, что ожидаемые и фактические объемы согласуются, считается верной. Для уровня значимости $\alpha = 0{,}95$ критической областью является интервал $(0{,}71; \infty)$, и, так как $\chi_r^2 = 9{,}12$ попадает в критическую область, нулевая гипотеза должна быть отклонена.

Рассмотрим следующую задачу:

Результаты исследования числа покупателей в аптеке в зависимости от времени работы приведены ниже:

Часы работы	9-10	10-11	11-12	12-13
Число покупателей	41	82	117	72

Возьмем среднее значение из промежутка:

m_i	9,5	10,5	12,5	11,5
x_i	41	82	72	117

Определить, починяется ли случайная величина X – число посетителей – нормальному закону?

По выборочным данным получим оценки параметров нормального закона распределения по формулам 5-6:

$$\bar{x} = \frac{\sum_{i=1}^{4} m_i x_i}{\sum_{i=1}^{4} m_i} = 79{,}45 \qquad (5)$$

$$s^2 = \frac{n}{n-1} d_B = \frac{4}{3}\left(\overline{x^2} - (\bar{x})^2\right) = \frac{4}{3}(7018{,}07 - 6312{,}30) = 941{,}03; s = 30{,}68(6)$$

Где, $\bar{x}$-выборочная средняя, s^2-исправленная выборочная дисперсия, n-объем выборки, d_B-выборочная дисперсия.

Находим значения случайной величины Z

$$z_i = \frac{x_i - \bar{x}}{s}, \quad (7)$$

По нормированным значениям величины Z находим значения функции Лапласа $\Phi(z)$.

Воспользуемся табличными значениями функции Лапласа $\Phi(z)$, далее найдем значение функции о распределении по формуле 8:

$$F_N(x_i) = 0{,}5 + \Phi(z), \quad (8)$$

Соберем все данные в таблицу 5.

Таблица 5

Расчет значений

m_i	0	9,5	10,5	12,5	11,5
z_i	0,00	-1,25	0,08	-0,24	1,22
$\Phi(z_i)$	-0,5000	-0,3944	0,0319	-0,0948	0,3883
$F_N(x_i)$	0,0000	0,1056	0,5319	0,4052	0,8883
$F_N(x_{i+1})$	0,1056	0,5319	0,4052	0,8883	1,000
$p_i^T = F_N(x_{i+1}) - F_N(x_i)$	0,1056	0,4263	-0,1267	0,4831	0,1117
$m_i^T = p_i^T n$	0,4224	1,7052	-0,5068	1,9324	0,4468
$\frac{(m_i - m_i^T)^2}{m_i^T}$	35,63154	-239,048	57,7904	273,4405	35,63154

По формуле 1 рассчитаем величину критерия согласия:

$$\chi_r^2 = \sum_{i=1}^{5} \frac{\left(m_i - m_i^T\right)^2}{m_i^T} = 127{,}8142$$

По таблице χ^2 для $\alpha = 0{,}05$ и числа степеней свободы $l = k - r - 1$=4-2-1=1 $\chi^2 = 3{,}8$. Следовательно, критическая область $(3{,}8; +\infty)$.

Величина χ_r^2 входит в критическую область, поэтому гипотеза о том, что случайная величина X – число посетителей – подчинена нормальному закону распределения, не согласуется с выборочными данными.

Мы выяснили, что проверка гипотез основывается на определении значения критерия согласия Пирсона и последующем анализе его принадлежности к

критической области. В случае вхождения найденного значения в критическую область нулевая гипотеза отвергается, в ином случае, она принимается верной.

Проверка гипотез о распределении по критерию Пирсона представляет собой сложную задачу, требующую хороших теоретических знаний в области статистики, но существование определенного алгоритма проверки гипотез значительно облегчает эту задачу.

Проверка гипотез о распределении случайной величины используется в разных сферах нашей жизни и потому является очень важным разделом математической статистики.

References

1. Карасев В.А. Статистика. Проверка гипотезы о виде закона распределения. / В.А. Карасев. – 3–е изд., – Москва: МИСИС, 2017. – 56 с. – ISN 978–5–906846–83–9. – Текст: непосредственный.

2. Максимов Ю.Д. Высшая математика/ Ю.Д. Максимов. – 2–е изд., – Москва: Проспект, 2019. – 327 с. – ISN 978–5–392–16271–0. – Текст: непосредственный.

3. Положинцев Б.И. Теория вероятностей и математическая статистика/ Б.И. Положинцев. – 4–е изд., – Санкт–Петербург: СПбПУ, 2016. – 95 с. – ISN978–5–7422–6083–7. – Текст: непосредственный

4. Губарь Л.Н. Теория вероятностей и математическая статистика/ Л.Н.Губарь. – 3–е изд., –Сыктывкар: Издательство СГУ имени Питирима Сорокина, 2015. – 120 с. – ISN 978–5–906810–13–7. – Текст: непосредственный